MULTI-CHANNEL SPEECH ENHANCEMENT BASED ON INDEPENDENT VECTOR EXTRACTION

Jaroslav Čmejla and Zbyněk Koldovský

Acoustic Signal Analysis and Processing Group,
Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Liberec, Czech Republic

ABSTRACT

In this paper, a gradient-based algorithm for Independent Vector Extraction is introduced and is first used for blind audio source extraction. We also propose novel modifications of the gradient learning rule in the algorithm based on preconditioning of input data and by using the AdaGrad update. Experiments with six-channel CHiME-4 recordings are conducted where the modified algorithms show significant improvement in terms of convergence speed.

Index Terms— Blind Audio Source Separation, Speech Enhancement, Independent Component Analysis, Independent Vector Analysis, Independent Vector Extraction

1. INTRODUCTION

The idea to separate speech signals using their statistical independence has become very popular after efficient methods for Independent Component Analysis (ICA) were developed [1]. In a time-frequency domain, a mixture of d original signals observed by d microphones can be, within the kth frequency bin, approximated by the instantaneous mixing model

$$\mathbf{X}^k = \mathbf{A}^k\mathbf{S}^k, \qquad k = 1, \dots, K, \tag{1}$$

where $\mathbf{S}^k$ and $\mathbf{X}^k$ denote $d \times N$ matrices of the original and mixed signals, respectively, where the rows and columns correspond to signals and to frames (time), respectively; $\mathbf{A}^k$ is the $d \times d$ nonsingular mixing matrix.

In the classical Frequency-Domain ICA approach [2], ICA is applied to each mixture separately, which gives rise to the permutation problem (the order of separated frequency components is random). To achieve the separation in the time-domain, a subsequent grouping of the frequency components has to be done [3]. Also, the scales of the separated components are random, which can be efficiently resolved by estimating spatial source images. Least-squares projections or the inverse matrix of the estimated $\mathbf{A}^k$ can be used for this; a recent study of these techniques appears in [4].

Independent Vector Analysis (IVA) was proposed to separate all signal mixtures in (1) jointly, which provides an alternative way for solving the permutation problem. The signals are separated as independent vector components whose elements should be as dependent as possible [5]. The approach has become popular, and many recent blind audio source separation methods follow the idea [6].

This work was supported by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-18-1-2040, and by the Student Grant Scheme 2018 project of the Technical University in Liberec.

In many speaker enhancement applications, only one particular speaker (from here referred to as the source of interest - SOI) should be focused using an array of microphones. ICA/IVA can be used to solve this task. However, since they separate that many sources as is the number of microphones d, they are computationally expensive once $d \gg 2$. Methods of Blind Source Extraction (BSE) appears to be more suitable in this case.

The BSE problem has been studied even earlier than ICA [7], and there are many blind extraction methods, e.g., based on maximizing the non-Gaussianity of the output signal [8] and are closely related to ICA; see, e.g., [9]. The practical problem in BSE is to ensure that the SOI is being extracted, not a different source. This cannot be guaranteed without additional information. Typically, an initial guess in the form of approximate direction of arrival (DOA) is given, nevertheless, algorithms need to be controlled during the computation so as not to change the focused source. The convergence can be stabilized by imposing constraints or by adding penalty terms to the objective function; see, e.g., [10, 11]. Also, pilot components can be used, if available; see, e.g., [12].

In this paper, we revise this problem as follows. A recently developed algorithm designed for extracting one independent vector component (so-called Independent Vector Extraction - IVE) is introduced. We apply this algorithm for blind audio source extraction and propose its two modifications: (a) a pre-conditioned version that enables us to control the SOI attraction area using known input signal-to-interference ratio, and (b) a modification of the gradient learning rule using approaches borrowed from the neural network learning theory. We show how these modifications can be used for increasing the speed of convergence of the algorithm. We also show how our method works in comparison to Independent Component Analysis (IVA).

The IVE problem formulation and the OGIVE algorithm are briefly described in Section 2. The proposed modifications are derived in Section 3 and experimentally validated in Section 4. Section 5 concludes the paper.

2. INDEPENDENT VECTOR EXTRACTION

2.1. Problem Statement

In IVE, the mixing matrices $\mathbf{A}^k$, $k = 1, \dots, K$, are parameterized to extract only one source in $\mathbf{S}^k$; let the SOI be the first source (row) in $\mathbf{S}^k$, denoted by the row vector $\mathbf{s}^k$. Let $\mathbf{A}^k$ be partitioned as $\mathbf{A}^k = [\mathbf{a}^k, \mathbf{Q}^k]$, the inverse matrix $(\mathbf{A}^k)^{-1} = \mathbf{W}^k$ be partitioned as $\mathbf{W}^k = [(\mathbf{w}^k)^H; \mathbf{B}^k]$, and $\mathbf{a}^k = [\gamma^k; \mathbf{g}^k]$. Since $\mathbf{B}^k$ is only required to be orthogonal to $\mathbf{a}^k$, we can select its concrete form $\mathbf{B}^k = [\mathbf{g}^k \quad -\gamma^k\mathbf{I}_{d-1}]$ where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix.

 978-1-5386-8151-0/18/$31.00

Let $\mathbf{w}^k = [\beta^k;\, \mathbf{h}^k]$. From $(\mathbf{A}^k)^{-1} = \mathbf{W}^k$ it follows that

$$\mathbf{W}^k = \begin{pmatrix} (\mathbf{w}^k)^H \\ \mathbf{B}^k \end{pmatrix} = \begin{pmatrix} (\beta^k)^* & (\mathbf{h}^k)^H \\ \mathbf{g}^k & -\gamma^k \mathbf{I}_{d-1} \end{pmatrix}, \tag{2}$$

and

$$\mathbf{A}^k = [\mathbf{a}^k, \quad \mathbf{Q}^k] = \begin{pmatrix} \gamma^k & (\mathbf{h}^k)^H \\ \mathbf{g}^k & \frac{1}{\gamma^k}\left(\mathbf{g}^k(\mathbf{h}^k)^H - \mathbf{I}_{d-1}\right) \end{pmatrix}, \tag{3}$$

where β^k and γ^k satisfy $\beta^k\gamma^k = 1 - (\mathbf{h}^k)^H\mathbf{g}^k$. Now, the re-parameterized mixing model can be written as

$$\mathbf{X}^k = \mathbf{A}^k\mathbf{V}^k, \tag{4}$$

where $\mathbf{V}^k = [\mathbf{s}^k; \mathbf{Z}^k]$, and $\mathbf{Z}^k = \mathbf{B}^k\mathbf{X}^k$.

Although the mixture model for $k = 1, \ldots, K$ are algebraically independent, a joint statistical model can be considered as in IVA [5]. Let s^k, $\mathbf{z}^k$, $\mathbf{v}^k$ symbolize the (vector) random variables having the same distributions as the samples in $\mathbf{s}^k$, $\mathbf{Z}^k$, $\mathbf{V}^k$, respectively, and let $\mathbb{s} = [s^1; \ldots; s^K]^T$, $\mathbb{z} = [\mathbf{z}^1; \ldots; \mathbf{z}^K]$, and $\mathbb{v} = [\mathbf{v}^1; \ldots; \mathbf{v}^K]$. By assuming the independence of $\mathbb{s}$ and $\mathbb{z}$, the joint pdf of $\mathbb{v}$ reads

$$p_{\mathbb{v}}(\mathbb{v}) = p_{\mathbb{s}}(\mathbb{s})p_{\mathbb{z}}(\mathbb{z}), \tag{5}$$

which means that $\mathbb{s}$ and $\mathbb{z}$ are independent but the elements inside of them can be dependent. Next, we assume that $\mathbb{z}$ are circular Gaussian with covariance matrix $\mathbf{C}_{\mathbb{z}} = \mathrm{E}[\mathbb{z}\mathbb{z}^H]$. We constrain our attention to uncorrelated mixtures $\mathbf{X}^1, \ldots, \mathbf{X}^K$, which means that $\mathbf{C}_{\mathbb{z}}$ is block-diagonal where the kth block is $\mathbf{C}_{\mathbf{z}}^k = \mathrm{E}[\mathbf{z}^k(\mathbf{z}^k)^H]$. Uncorrelated mixtures arise, for instance, in the frequency-domain separation [5, 13].

Now, the quasi-log-likelihood contrast function (for one signal sample) for the estimation of parameter vectors $\mathbf{a}^k$ and $\mathbf{w}^k$, $k = 1, \ldots, K$, is given by

$$\mathcal{J}(\mathbf{a}^1, \ldots, \mathbf{a}^K, \mathbf{w}^1, \ldots, \mathbf{w}^K) = \log f\big(\widehat{s}^1, \ldots, \widehat{s}^K\big) - \sum_{k=1}^{K} \mathbf{x}^{k\,H}\mathbf{B}^{k\,H}(\mathbf{C}_{\mathbf{z}}^k)^{-1}\mathbf{B}^k\mathbf{x}^k + \sum_{k=1}^{K}\log|\det\mathbf{W}^k|^2, \tag{6}$$

where $\widehat{s}^k = (\mathbf{w}^k)^H\mathbf{x}^k$, and $f(\cdot)$ is the model pdf replacing the unknown density $p_{\mathbb{s}}(\cdot)$.

2.2. Orthogonally Constrained Gradient Algorithm

We now describe an algorithm for finding the local maximum of (6) that performs small-step updates in the direction of the gradient of (6) with respect to $\mathbf{w}^k$ when $\mathbf{a}^k$ is dependent on $\mathbf{w}^k$ through

$$\mathbf{a}^k = \frac{\widehat{\mathbf{C}}_{\mathbf{x}}^k\mathbf{w}^k}{\mathbf{w}^{k\,H}\widehat{\mathbf{C}}_{\mathbf{x}}^k\mathbf{w}^k}, \quad k = 1, \ldots, K, \tag{7}$$

where $\widehat{\mathbf{C}}_{\mathbf{x}}^k = \mathbf{X}(\mathbf{X}^k)^H/N$ is the sample covariance matrix of $\mathbf{X}^k$. The constraint (7) is imposed in order to avoid spurious maxima of (6) where $\mathbf{w}^k$ and $\mathbf{a}^k$ do not correspond to the same source [14].

Since $\mathbf{C}_{\mathbf{z}}^k$ is not known, it is replaced in (6) by its current sample-based value $\widehat{\mathbf{C}}_{\mathbf{z}}^k = \mathbf{Z}^k(\mathbf{Z}^k)^H/N$. Then, the gradient of (6) is equal to

$$\frac{\partial\mathcal{J}}{\partial\mathbf{w}^{k\,H}} = \mathbf{a}^k - \frac{1}{N}\mathbf{X}^k\phi^k\big(\widehat{\mathbf{s}}^1, \ldots, \widehat{\mathbf{s}}^K\big)^T, \tag{8}$$

where $\widehat{\mathbf{s}}^k = (\mathbf{w}^k)^H\mathbf{X}^k$, and

$$\phi^k(\xi^1, \ldots, \xi^K) = -\frac{\partial\log f(\xi^1, \ldots, \xi^K)}{\partial\xi^k}, \tag{9}$$

is the kth score function related to the model density $f(\cdot)$. By a simple inspection of (8) for $N \to +\infty$, it can be shown that the true extraction vector $\mathbf{w}^k$ is the stationary point of (8) if and only if $\mathrm{E}[\phi^k(s^1, \ldots, s^K)s^k] = 1$. This property is satisfied by the true score function, which is, however, not known in the blind scenario. Therefore, $\phi^k(\cdot)$ is, in every step, normalized so that the average value of $\phi^k(\widehat{s}^1, \ldots, \widehat{s}^K)\widehat{s}^k$ is equal to one (see lines 11 and 12 in Algorithm 1).

The gradient algorithm, from here referred to as OGIVE, starts from initial values and, in each iteration, it updates $\mathbf{w}_k$ by adding a "small" step in the direction of the gradient (8) for each $k = 1, \ldots, K$; for illustration, OGIVE is summarized in Algorithm 1; for more details, see [15].

Algorithm 1: OGIVE: orthogonally constrained extraction of an independent vector component from the set of mutually uncorrelated mixtures $\mathbf{X}^1, \ldots, \mathbf{X}^K$

Input: $\mathbf{X}^k$, $\mathbf{w}_{\mathrm{ini}}^k$, $k = 1, \ldots, K$, μ, `tol`
Output: $\mathbf{a}^k$, $\mathbf{w}^k$, $k = 1, \ldots, K$

1 **foreach** $k = 1, \ldots, K$ **do**
2 $\quad\widehat{\mathbf{C}}_{\mathbf{x}}^k = \mathbf{X}^k(\mathbf{X}^k)^H/N$;
3 $\quad\mathbf{w}^k = \mathbf{w}_{\mathrm{ini}}^k$;
4 **end**
5 **repeat**
6 $\quad$**foreach** $k = 1, \ldots, K$ **do**
7 $\quad\quad\mathbf{a}^k \leftarrow ((\mathbf{w}^k)^H\widehat{\mathbf{C}}_{\mathbf{x}}^k\mathbf{w}^k)^{-1}\widehat{\mathbf{C}}_{\mathbf{x}}^k\mathbf{w}^k$;
8 $\quad\quad\widehat{\mathbf{s}}^k \leftarrow (\mathbf{w}^k)^H\mathbf{X}^k$;
9 $\quad$**end**
10 $\quad$**foreach** $k = 1, \ldots, K$ **do**
11 $\quad\quad\nu^k \leftarrow \widehat{\mathbf{s}}^k\phi^k(\widehat{\mathbf{s}}^1, \ldots, \widehat{\mathbf{s}}^K)^T/N$;
12 $\quad\quad\Delta^k \leftarrow \mathbf{a}^k - (\nu^k)^{-1}\mathbf{X}^k\phi^k(\widehat{\mathbf{s}}^1, \ldots, \widehat{\mathbf{s}}^K)^T/N$;
13 $\quad\quad\mathbf{w}^k \leftarrow \mathbf{w}^k + \mu\Delta^k$;
14 $\quad$**end**
15 **until** $\max\{\|\Delta^1\|, \ldots, \|\Delta^K\|\} < \mathtt{tol}$;

3. PROPOSED MODIFICATIONS

3.1. Preconditioning

Owing to the multiplicative form of the models (1) and (4), each gradient update can be considered subject to input data that were multiplied by a preconditioning matrix $\mathbf{D}^k$, i.e., subject to "new" data $\mathbf{U}^k = \mathbf{D}^k\mathbf{X}^k$. For simplicity, we can omit the index k and assume $K = 1$ for now.

Let $\mathbf{w}_{\mathbf{x}}$ and $\mathbf{w}_{\mathbf{u}}$ be the separating vectors operating on $\mathbf{X}$ and $\mathbf{U}$, respectively, which give the same output, that is, $\widehat{\mathbf{s}} = \mathbf{w}_{\mathbf{x}}^H\mathbf{X} = \mathbf{w}_{\mathbf{u}}^H\mathbf{U}$. It follows that $\mathbf{w}_{\mathbf{x}} = \mathbf{D}^H\mathbf{w}_{\mathbf{u}}$. The sample covariance matrix of $\mathbf{U}$ is given by $\widehat{\mathbf{C}}_{\mathbf{u}} = \mathbf{D}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{D}^H$.

Consider now the gradient of (6) when the input data are $\mathbf{U}$ and the starting vector is $\mathbf{w}_{\mathbf{u}}$. The gradient given by the right-hand side

of (8) with the constraint (7) then reads

$$\Delta_{\mathbf{u}} = \frac{\widehat{\mathbf{C}}_{\mathbf{u}}\mathbf{w}_{\mathbf{u}}}{\mathbf{w}_{\mathbf{u}}^H\widehat{\mathbf{C}}_{\mathbf{u}}\mathbf{w}_{\mathbf{u}}} - \frac{1}{N}\mathbf{U}\phi\big(\widehat{\mathbf{s}}^1,\ldots,\widehat{\mathbf{s}}^K\big)^T = \frac{\mathbf{D}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{D}^H\mathbf{w}_{\mathbf{u}}}{\mathbf{w}_{\mathbf{u}}^H\mathbf{D}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{D}^H\mathbf{w}_{\mathbf{u}}} - \frac{1}{N}\mathbf{D}\mathbf{X}\phi\big(\widehat{\mathbf{s}}^1,\ldots,\widehat{\mathbf{s}}^K\big)^T = \mathbf{D}\left(\frac{\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{w}_{\mathbf{x}}}{\mathbf{w}_{\mathbf{x}}^H\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{w}_{\mathbf{x}}} - \frac{1}{N}\mathbf{X}\phi\big(\widehat{\mathbf{s}}^1,\ldots,\widehat{\mathbf{s}}^K\big)^T\right) = \mathbf{D}\Delta_{\mathbf{x}} \quad (10)$$

Then, the updated vector $\mathbf{w_u}$ applied to the data $\mathbf{U}$ gives

$$(\mathbf{w}_{\mathbf{u}}^{\text{new}})^H\mathbf{U} = (\mathbf{w}_{\mathbf{u}} + \mu\Delta_{\mathbf{u}})^H\mathbf{U} = \left(\mathbf{D}^{-H}\mathbf{w}_{\mathbf{x}} + \mu\mathbf{D}\Delta_{\mathbf{x}}\right)^H\mathbf{D}\mathbf{X} = \left(\mathbf{w}_{\mathbf{x}} + \mu\mathbf{D}^H\mathbf{D}\Delta_{\mathbf{x}}\right)^H\mathbf{X}. \quad (11)$$

This gives us a generalized update rule for $\mathbf{w_x}$

$$\mathbf{w}_{\mathbf{x}} \leftarrow \mathbf{w}_{\mathbf{x}} + \mu\mathbf{D}^H\mathbf{D}\Delta_{\mathbf{x}}, \quad (12)$$

which coincides with the original update rule (line 13 in Algorithm 1) when $\mathbf{D} = \mathbf{I}_d$.

There are special choices of $\mathbf{D}$ already known in the literature. For $\mathbf{D} = \mathbf{W}$, (12) gives an analogy of the widely known *natural* gradient for ICA [16, 17]. This choice corresponds to the update when the input data are pre-separated by the current de-mixing matrix prior to each iteration while the starting $\mathbf{w}$ is equal to the unit vector.

Another choice is $\mathbf{D} = \mathbf{R}\widehat{\mathbf{C}}_{\mathbf{x}}^{-1/2}$, where $\widehat{\mathbf{C}}_{\mathbf{x}}^{-1/2}$ denotes the inverse of the matrix square root of $\widehat{\mathbf{C}}_{\mathbf{x}}$ and $\mathbf{R}$ is an arbitrary unitary matrix; here, it holds that $\mathbf{D}^H\mathbf{D} = \widehat{\mathbf{C}}_{\mathbf{x}}^{-1}$. This choice corresponds to the update realized on whitened (uncorrelated and normalized) signals [18].

The preconditioning matrix has an influence on the surface of the contrast function. When all source should be separated, as in ICA, the widths of convergence areas of particular sources, influenced by that matrix, appear to be unimportant because all sources are separated. However, when only one source should be extracted, the width of the area of convergence to the SOI becomes essential since it has an influence on the success of the extraction.

Our previous experimental studies [15] have shown that the attraction areas are mainly influenced by an initial Signal-to-Interference Ratio (SIR_{in})[1]. However, SIR_{in} is not defined uniquely when there are multiple channels. Nevertheless, we can restrict our attention to special cases where SIR_{in} is approximately the same on all channels. Then, it is meaningful to define SIR_{in}, e.g., as its average value taken over all channels. For example, this happens in cases of miniature microphone arrays where the distance of sources are much larger that the distances of microphones.

There are two important observations from our previous studies. First, when there is a significantly dominant source in the mixture, the update rule (12) with $\mathbf{D} = \mathbf{I}_d$ tends to converge to it from almost any initial point. Second, when there is a source having a significantly smaller SIR_{in} compared to the other sources, (12) tends to converge to it with $\mathbf{D} = \widehat{\mathbf{C}}_{\mathbf{x}}^{-1/2}$. Based on these two observations, we propose a modification of OGIVE such that the preconditioning matrix $\mathbf{D}$ is selected based on SIR_{in} (related to the SOI) and the update rule (line 13 in Algorithm 1) is performed according to (12).

More specifically, consider $K > 1$, so $\mathbf{D}^k$ depends on k. Typically, SIR_{in} depends on k as well since it is varying across frequencies[2]. Since SIR_{in} is not known in the blind scenario, we consider two solutions: An "oracle" modification of OGIVE puts $\mathbf{D}^k = \mathbf{I}_d$ when SIR_{in} for the kth frequency is higher than 5 dB, otherwise $\mathbf{D}^k = (\widehat{\mathbf{C}}_{\mathbf{x}}^k)^{-1/2}$. The other solution ("precond"), which does not require prior knowledge, puts $\mathbf{D}^k = \mathbf{I}_d$ for k corresponding to frequencies 156 Hz through 2.5 kHz[3]; otherwise $\mathbf{D}^k = (\widehat{\mathbf{C}}_{\mathbf{x}}^k)^{-1/2}$.

In OGIVE, the parallel extraction sub-algorithms (one for each frequency) influence each other. We therefore cannot simply expect that the above proposed modifications would guarantee convergence of each sub-algorithm. However, these modification can significantly improve the overall convergence speed of OGIVE, which is verified by experiments in Section 4.

3.2. AdaGrad Learning-rule Update

A difficulty in first-order optimization methods is the proper setting of the learning rate parameter. AdaGrad [19] provides an approach trying to avoid this by learning rate adaptation over iterations. Recursively adapted learning rate is given by square root of the sum of the previous gradient. In our case, this means that each element of $\mathbf{w}^k$ can have a different learning rate.

The AdaGrad update rule is defined through

$$\mathbf{q}_{t+1}^k = \mathbf{q}_t^k + (\Delta_t^k \odot \Delta_t^k) \quad (13)$$

$$\mathbf{w}_{t+1}^k \leftarrow \mathbf{w}_t^k + (\mu\Delta_t^k) \oslash (\sqrt{\mathbf{q}_{t+1}^k} + \epsilon) \quad (14)$$

where t denotes iteration index, ϵ denotes a small constant that prevents from division by zero (our choice is $\epsilon = 10^{-7}$), and $\mathbf{q}_t^k$ is a vector of the same length as $\mathbf{w}_t^k$ that accumulates squared gradients over the iterations; $\mathbf{q}_0^k$ is set to the zero vector. Operators $\odot$ and $\oslash$ denotes element-wise multiplication and division, respectively.

4. EXPERIMENTS

4.1. Settings

We conducted an experimental validation using the 4$^{\text{th}}$ CHiME simulated development dataset [20]. This dataset contains recordings of speech signals having about 4 to 12 seconds in length and interference signals from various environments: bus, coffee restaurant, pedestrian area and street. The recordings were acquired using an array of six microphones mounted to a tablet; the sampling frequency is 16 kHz. In total, there are 2960 mixtures of speech with environmental noise; 740 mixtures per environment.

Speech and noise signals were mixed so that the average signal-to-noise ratio (SNR) over all microphones was equal to 0 dB. Signals were processed in the STFT domain with the frame length of 512 samples and 75% overlap. With assumption that speaker was right in front of the tablet (perpendicular position to the microphone array), the initial mixing vector for the OGIVE was set to be vector of ones for all frequency bins. The same initialization was used for the first source of the IVA. Initialization for the rest of sources was set to be orthogonal to the first source (same as it is described in 2.1).

[1] It is worth pointing out here that SIR_{in} does not influence the final separation accuracy as the ICA problem is equivariant [1]. In other words, the preconditioning matrix does not influence the optimum points of the contrast function. However, it influences its global surface, which affects the convergence trajectory of the algorithm.

[2] For example, when the SOI is a speaker, SIR_{in} can take values from -20 to 20 dB.

[3] This range was selected ad hoc according to a typical dominant band of speech signals.

Table 1. Statistics over all mixtures in dataset (SNR improvement after 1000 iterations).

	IVE						IVA
	Normal			AdaGrad			Scaled Natural Gradient
	no precond	precond	oracle precond	no precond	precond	oracle precond	
mean	9.71	11.43	10.93	10.21	9.89	9.83	11.40
median	9.93	11.46	11.25	10.13	10.19	10.11	11.95
std	4.47	4.36	5.65	3.85	4.65	5.42	5.34
min	-13.29	-15.55	-25.82	-13.61	-26.64	-22.99	-24.74
max	39.36	38.14	37.81	27.21	26.58	36.64	31.44
$\mathbf{SNR_{imp} < -10dB}$	0.17%	0.24%	1.15%	0.14%	0.44%	1.28%	1.18%
$\mathbf{SNR_{imp} < 0dB}$	2.06%	1.05%	3.28%	0.71%	2.87%	4.05%	2.91%

This ensures that the source of the interest appears on the first output channel of the IVA. For IVA the Scaled Natural Gradient [21] was used to obtain better stability and faster convergence.

Since there is the scaling ambiguity in the IVE/IVA algorithms, the spatial image of the extracted source on the first microphone was computed using the Minimal Distortion Principle [22]. The extraction performance was evaluated by the SNR improvement.

Fig. 1 shows results achieved by combinations of the proposed variants of OGIVE in Sections 3.1 and 3.2 in terms of median improvement of SNR as it evolves over iterations. The median was used due to large variability of the mixtures; detailed statistics are given in Table 1. Since the optimum learning rate μ is dependent on the variant of OGIVE/IVA, we selected values that yield optimum results in the final evaluation. Here, it worthy noting that the AdaGrad modification appeared to be less sensitive to this choice.

The results in Fig. 1 show 9-12 dB of median SNR improvement by all variants of OGIVE after 1000 iterations. The methods endowed by the oracle preconditioning show the highest convergence speed, which confirms our expectations described in Section 3.1 that the preconditioning based on the initial SNR leads to the extension of the convergence area and to the increased speed of convergence. Also, the non-oracle "precond" rule leads to an increased speed of the convergence as compared to the variants without preconditioning. The IVA achieves better final results, but at the cost of slower convergence. This can be caused by the fact that in the OGIVE algorithm the background noise is modeled as Gaussian. On the other hand IVA considers the background to be non-Gaussian.

AdaGrad brings increased speed of convergence as compared to the normal update rule, however, only when no preconditioning is applied. In the other cases, AdaGrad performed comparably to the normal update rule.

The statistics in Table 1 show that there were cases where the SNR improvement by the methods was below -10 dB. This happens when a different source than the SOI is focused by the algorithm. The number of these cases is higher when preconditioning is used. This is probably an effect of the faster and better convergence, although to a different source.

5. CONCLUSION

In this paper, we have shown the capability of OGIVE to extract a speaker from 6-channel audio mixtures. We have proposed modifications in order to speed up the convergence of the algorithm through preconditioning and employing the AdaGrad update rule. Experiments evaluated on the CHiME-4 dataset demonstrate the performance of both modifications in terms of convergence speed. We found out that OGIVE, in combination with a correctly chosen preconditioning, shows an increased convergence speed for both versions of gradient update rule. The AdaGrad update shows faster convergence than the normal update rule only when there is no preconditioning. In our future work, we plan to focus on the application of these findings for on-line implementations of ICE/IVE algorithms, where the speed of convergence is very important for a practical deployment of methods in dynamic environments. Another goal is to achieve at least the same performance as the IVA.

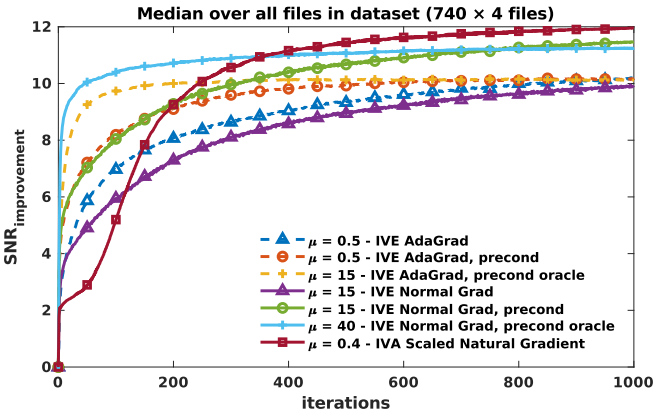

Fig. 1. Results in terms of median SNR improvement as functions of the number of iterations.

6. REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Independent Component Analysis and Applications Series. Elsevier Science, 2010.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[3] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[4] Zbyněk Koldovský and Francesco Nesta, "Performance analysis of source image estimators in blind source separation," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4166–4176, Aug. 2017.

[5] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 70–79, Jan. 2007.

[6] Nobutaka Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[7] Peter J. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, no. 2, pp. 435–475, June 1985.

[8] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, July 1997.

[9] J. F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct 1998.

[10] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, Sep 2002.

[11] Affan H. Khan, Maja Taseska, and Emanuël A. P. Habets, *A Geometrically Constrained Independent Vector Analysis Algorithm for Online Source Extraction*, pp. 396–403, Springer International Publishing, Cham, 2015.

[12] F. Nesta, S. Mosayyebpour, Z. Koldovský, and K. Paleček, "Audio/video supervised independent vector analysis through multimodal pilot dependent components," in *Proceedings of European Signal Processing Conference*, Sept. 2017, pp. 1190–1194.

[13] M. Anderson, G. S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, Sept 2014.

[14] Zbyněk Koldovský, Petr Tichavský, and Václav Kautský, "Orthogonally constrained independent component extraction: Blind MPDR beamforming," in *Proceedings of European Signal Processing Conference*, Sept. 2017, pp. 1195–1199.

[15] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction," Mar. 2018, arXiv:1803.10108 [eess.SP].

[16] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proceedings of Neural Information Processing Systems*, 1996, pp. 757–763.

[17] J. F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, Dec 1996.

[18] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[19] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, July 2011.

[20] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[21] S. C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, April 2007, vol. 2, pp. II–637–II–640.

[22] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, Dec. 2001, pp. 722–727.